#12

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 07-287969

(43)Date of publication of application : 31.10.1995

(51)Int.Cl. G11B 27/28
G10L 3/00
G10L 3/00
G10L 3/00
H04R 3/00

(21)Application number : 07-082900

(22)Date of filing : 07.04.1995

(71)Applicant : XEROX CORP

(72)Inventor : VIJAY BALASUBRAMANIAN
CHEN FRANCINE R
PHILIP A CHOU
DONALD G KIMBER
ALEX D POON
WEBER KARON A
LYNN D WILCOX

(30)Priority

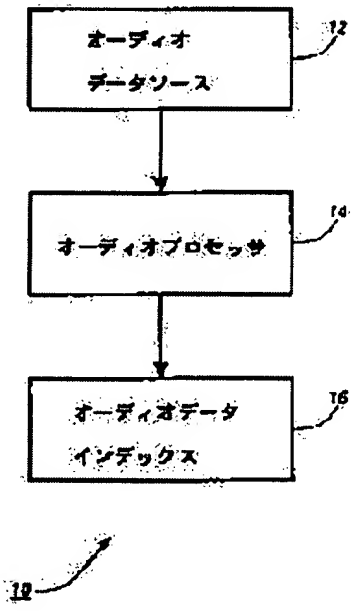
Priority number : 94 226580 Priority date : 12.04.1994 Priority country : US

(54) SYSTEM OF PROCESSOR CONTROL

(57)Abstract:

PURPOSE: To prepare an index in an audio data stream.

CONSTITUTION: An audio stream is given from an audio data source 12, and the data are imparted by a speaker conducting a conversation, a recording video with an audio track or other audio sources. Audio data are transmitted to an audio processor 14, the audio processor is arbitrarily known as a general purpose computer, and the audio processor outputs an audio data index 16.



LEGAL STATUS

[Date of request for examination] 08.04.2002

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号

特開平7－287969

(43)公開日　平成７年(1995)10月31日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ１１Ｂ　27/28		Ａ　8224－5Ｄ		
Ｇ１０Ｌ　3/00	５３１　Ｌ			
	５３５			
	５５１　Ｇ			
Ｈ０４Ｒ　3/00	３１０			

審査請求　未請求　請求項の数３　ＯＬ　（全 13 頁）

(21)出願番号　特願平7－82900

(22)出願日　平成７年(1995)４月７日

(31)優先権主張番号　２２６５８０

(32)優先日　1994年４月12日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　590000798
ゼロックス　コーポレイション
ＸＥＲＯＸ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国　ニューヨーク州　14644
ロチェスター　ゼロックス　スクエア
（番地なし）

(72)発明者　ヴィジャイ　パラスプラマニアン
アメリカ合衆国　ニュージャージー州
08544プリンストン　プリンストン　ユニバーシティー　オールド　グラジュエイト　カレッジ　ルーム　184

(74)代理人　弁理士　中島　淳　（外１名）

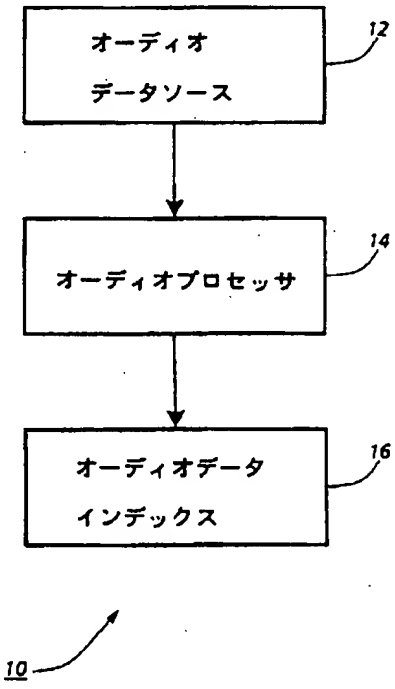
最終頁に続く

(54)【発明の名称】　プロセッサ制御のシステム

(57)【要約】

【目的】　オーディオデータストリーム内にインデックスを作成する。

【構成】　オーディオストリームはオーディオデータソース１２から与えられ、該データは、会話を行うスピーカー、オーディオトラックを伴う記録ビデオ、または他のオーディオソースによって与えられることが可能である。オーディオデータはオーディオプロセッサ１４へ送られ、オーディオプロセッサは汎用コンピュータのような任意の公知デバイスであることが可能であり、本発明に従って構成されることが可能である。オーディオプロセッサはオーディオデータインデックス１６を出力する。



1

【特許請求の範囲】

【請求項１】　リアルタイムに記録されるオーディオデータに対してスピーカーに従う電子インデックスを相関付けるプロセッサ制御のシステムであって、前記オーディオデータは複数の個々のスピーカーからの音声を含み、前記システムは、

複数の個々のスピーカーの各々に対するスペクトル特徴トレーニングデータを与えるトレーニングデータソースと、

トレーニングデータを受信して複数の個々のスピーカーの各々に対するスピーカーモデルを生成するシステムプロセッサであって、各スピーカーモデルは関連するスピーカー識別子を有し、前記システムプロセッサはさらに前記スピーカーモデルを結合してスピーカーネットワークとする、システムプロセッサと、

複数の個々のスピーカーからの音声を含むリアルタイムオーディオデータを与えるオーディオ入力システムと、

前記オーディオデータを受信して前記オーディオデータをスペクトル特徴データに変換するオーディオプロセッサと、

前記オーディオデータを受信して受信時間に従って記憶媒体上へ前記オーディオデータを記録する記録デバイスと、

データを記憶するメモリであって、メモリに記憶されるデータはプロセッサが実行する命令を示す命令データを含む、メモリと、

を含み、

前記システムプロセッサはさらに、メモリに記憶されたデータにアクセスし、

前記システムプロセッサは、命令実行の際に、前記スペクトル特徴データを前記オーディオプロセッサから受信し、前記スピーカーネットワークを使用して、異なる個々のスピーカーモデルに対応する前記オーディオデータのセグメントを決定し、

前記システムプロセッサはさらに、各セグメントの開始においてタイムスタンプを決定し、前記タイムスタンプは前記記憶媒体上の当該セグメントに対する受信時間に対応し、前記システムプロセッサは前記タイムスタンプを前記メモリに記憶し、

前記システムプロセッサはさらに、当該セグメントに対する前記記憶媒体のロケーションアドレスと共に、各セグメントに対する前記個々のスピーカーモデルの前記スピーカー識別子を前記メモリに記憶する、

プロセッサ制御のシステム。

【請求項２】　前記システムプロセッサは複数の個々のスピーカーの各々に対して個々のＨＭＭスピーカーモデルを生成する、請求項１に記載のシステム。

【請求項３】　前記システムプロセッサはさらに前記個々のＨＭＭスピーカーモデルを並列に結合してスピーカーネットワークＨＭＭを形成する、請求項２に記載のシステム。

2

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は記録またはリアルタイムオーディオデータストリームに対する会話音声のスピーカー（話し手）によるセグメンテーションに関する。

【０００２】より詳細には、本発明は既知のスピーカーを含むオーディオデータのリアルタイム記録期間中に会話音声を自動的にセグメンテーションするシステムに関する。

【０００３】

【従来の技術】オーディオおよびビデオ記録は、コンシューマグレード（消費者レベル）の記録装置の発展によって今や一般のものとなっている。後の再生のための過去の記録としてビジネスミーティング、講義、もしくはバースデーパーティーが記録されることは今や稀なことではない。不幸にして、オーディオおよびビデオ媒体の両者は、所望の記録部分にアクセスする際のアシストとなる外部またはオーディオ情報をほとんど与えない。書籍においては、巻頭の目次および巻末の索引によってインデックス化が与えられ、このインデックス化によって読者は複数の著者の確認および複数の著者の参照を容易に行うことが可能である。同様のインデックス化方法がオーディオストリームにおいて有用であり、ユーザーは特定のスピーカーの会話部分を確認することが可能となる。ほとんどのビデオ記録に関連する限られたデータ量は、見る者が確実におよび容易に所望の関心部分にアクセスするための充分な情報を与えない。このため見る者は記録内容を順に調べて所望の情報を検索しなければならない。

【０００４】例えばスピーカー（話し手）やトピック（主題）を示すノートのような、記録中に取られたノートが検索の補助となることが可能である。このようなノートは構造的アウトラインを与えるが、ビデオ媒体とノート媒体との間には直接的な相関がないため、ノートの内容を共にしたビデオ上の時刻の補完を強いられる。このことは、非相関媒体におけるイベントノートは通常イベントの継続時間を含まないという事実によって複雑化する。加えて、そのようなノート化またはインデックス化は非常に煩わしい。コンピュータシステムがイベント期間中のノート取得に使用されることが可能であり、該システムは同時に記録されるかまたは事前に記録される。キーボードを使用するテキストベースシステムがこの場合に使用されることが可能であるが、ほとんどの人はタイプするよりもかなり速く話すため、内容を記述するコンピュータ生成テキストラベルをリアルタイムで作成することは相当な努力を必要とする。

【０００５】

【発明が解決しようとする課題】オーディオストリームにおいて異なるスピーカーを示すスピーカーチェンジマ

ーカーは、異なるシーケンシャルデータへのランダムアクセスを可能とする。リアルタイム設定においては、そのようなオーディオセグメンテーションは、記録が行われている時にその記録の中へ有用なインデックスを作成する際の補助となり得る。各セグメントは１個人による発声を表す。同一のスピーカーによる発声は結合され、また同様に参照されてインデックスが形成される。会話におけるポーズまたは沈黙期間もまたオーディオインデックス形成において重要である。

【０００６】オーディオストリーム内にインデックスを作成することは、リアルタイムであっても処理後であっても、ユーザーが特定のオーディオデータセグメントを認識することを可能にする。例えばこのことは、ユーザーが記録を拾い読みして特定のスピーカーに対応するオーディオセグメントを選択したり、次のスピーカーへ記録を早送りすることを可能にする。加えて、スピーカーの順序を知ることは、会話または会話の内容に関する内容情報を与えることも可能である。

【０００７】

【課題を解決するための手段】本発明は、リアルタイムに記録されるオーディオデータに対してスピーカーに従う電子インデックスを相関付けるプロセッサ制御のシステムを与える。該システムは、複数の個々のスピーカーの各々に対するトレーニングデータソースを含む。オーディオ入力システムは、個々のスピーカーに対する音声を含むリアルタイムオーディオデータを与える。オーディオデータはオーディオプロセッサによってスペクトル特徴データに変換されると同時に記録デバイスによって記憶媒体上に記録される。システムプロセッサはトレーニングデータを受信して個々のスピーカーモデルを作成し、該モデルは並列に接続されてスピーカーネットワークが形成される。システムプロセッサは次にオーディオデータのスペクトル特徴データを受信し、スピーカーネットワークを使用して各スピーカーに対応するオーディオデータ内のセグメントを決定する。

【０００８】

【実施例】図１は一般化されたオーディオ処理システム１０のブロック図を示し、該システムにおいて本発明が実施されることが可能である。一般に、オーディオストリームはオーディオデータソース１２から与えられ、該データは、会話を行うスピーカー、オーディオトラックを伴う記録ビデオ、または他のオーディオソースによって与えられることが可能である。オーディオデータはオーディオプロセッサ１４へ送られ、オーディオプロセッサは汎用コンピュータのような任意の公知デバイスであることが可能であり、本発明に従って構成されることが可能である。オーディオプロセッサはオーディオデータインデックス１６を出力する。

【０００９】図２はオーディオインデックスシステムの一般化されたフロー図を示す。図２に示されるステップは以下により詳細に説明されるが、図２は本発明により記述される方法の概観を与えるものである。

【００１０】オーディオ波形２０はボックス２２のステップにおける入力である。ボックス２２におけるオーディオストリームは、処理されるべきオーディオの部分を含むことが可能であるが、オーディオストリーム内の全てのスピーカーからの音声を含まなければならない。説明を目的として、オーディオストリーム全体がボックス２２のステップにおける入力である。ボックス２４のステップは音声信号データをスペクトル特徴ベクトルへ変換する。例えば、１２次のケプストラムが２０ｍｓごとに算出されることが可能である。

【００１１】ボックス２６のステップにおいて、ＨＭＭスピーカーモデルは初期化データに基づき各スピーカーに対してトレーニングされる。複数の個々のスピーカーモデルは、モデルを並列に接続することによってボックス２８のステップにおいて結合され、会話のＨＭＭスピーカーモデルが形成される。

【００１２】ボックス３０のステップは、セグメンテーションが実行されるオーディオストリームを入力する。オーディオストリームはボックス２２のステップで使用されるトレーニングオーディオデータを含んでも含まなくてもよい。スピーカーモデルの事前トレーニングに対してスピーカーが使用可能である場合、入力されるオーディオストリームもまたリアルタイムに発生およびセグメンテーションされることが可能である。ボックス３２のステップにおいて、入力されるオーディオから再び特徴が抽出され、この特徴抽出はボックス２４のステップにおけるものと同様である。

【００１３】ボックス３４のステップはボックス２８のＨＭＭスピーカーネットワークを使用し、入力されるオーディオストリームのセグメンテーションを行う。セグメンテーションはビタビ（Viterbi）デコーディングを使用して行われ、スピーカーネットワークを介する最も確からしい状態シーケンスが見出され、状態パスがスピーカーを変更する場合にはマーキングが施される。

【００１４】セグメンテーションとインデックス化の確度は、ボックス２６のステップに戻ってスピーカーモデルを再トレーニングすることによる後処理の適用で改善されることが可能であり、この場合ボックス３４のステップからのセグメンテーション情報が使用される。一般に、セグメンテーション情報に基づき、スピーカー当たりより多くの音声情報が使用可能となり、このことはより詳細なスピーカーモデルが決定されることを可能とする。再トレーニングと再セグメンテーションの繰り返しは、ボックス３４のステップでのセグメンテーションで大きな変化が生じなくなるまで続けられることが可能である。ボックス３２のステップにおける特徴抽出の結果はまたセーブされることが可能であり、各再トレーニングの繰り返しと共に再使用されてボックス３４のステッ

5

プでオーディオデータが再セグメンテーションされることが可能である。

【００１５】隠れマルコフモデル（HMM）によるモデル化は音声認識で一般的に使用される統計的方法であり、ワード全体、もしくは単音のようなサブワードがモデル化される。未知の発声の認識は、その発声が最も確からしく与えられるモデルもしくはモデルのシーケンスを見出すことに基づいている。HMMはスピーカーの識別においても使用されることが可能である。モデルはスピーカーの発音に対して作成され、その場合発音は特定のワードについてのものであっても自然な音声についてのものであってもよい。スピーカーの識別は、未知の発声が最も確からしく与えられるスピーカーモデルを見出すことによって行われる。未知の発声が複数のスピーカーからの音声を含む場合、スピーカーは最も確からしいスピーカーモデルのシーケンスを見出すことによって識別される。

【００１６】理論的に、HMMは状態のシーケンスから成り、該状態シーケンスは定められた時間間隔で状態間に発生する遷移を伴う。ある状態への遷移が行われるたびに、その状態の出力特性が発生される。音声認識およびスピーカー識別の両者において、これらの出力はその時間間隔に対する音声のスペクトル推定を表す。例えばケプストラムがその例である。ケプストラムはスペクトルエンベロープ（包絡線）の推定であり、音声認識およびスピーカー識別で一般に使用される。ケプストラムは、スペクトルの対数のフーリエ逆変換であり、スペクトルエンベロープと周期的音声ソースとを分離するよう作用する。

【００１７】状態間の遷移は出力のシーケンスを特定する。状態間遷移および各状態出力に確率を関連付けることによって、HMMが使用されて音声を統計的にモデル化することが可能となる。システムの出力のみが観測されるため「隠れ（hidden）」という言葉が用いられる。即ち、基礎となる状態シーケンスは推定され得るのみである。

【００１８】より形式的には、HMM　L　は、$S_0 \ldots S_{N-1}$ のN個の状態、状態iから状態jへの遷移確率 $a_{ij}, i=0 \ldots N-1, j=0 \ldots N-1$、ならびに状態iで出力xを生じる確率を与える確率分布 $b_i(x), i=0 \ldots N-1$、から成る。例えば、$b_i(x)$ は特徴ベクトル xに対する多変数ガウス分布であることが可能である。加えて、遷移可能であるが出力を発生しないヌル状態が存在する。図３は５状態のHMMを示す。状態S_0 から状態S_1、S_2 またはS_3 への遷移確率は画一的であり、即ち、$a_{0j}=1/3, j=1,2,3$である。状態$S_i, i=1,2,3$については、自己遷移および状態S_4 への遷移が存在し、それらは等確率である。従って$a_{11}=1/2$および$a_{14}=1/2, i=1,2,3$である。状態S_4 については遷移は常にS_0 へ行われ、従って$a_{40}=1$である。状態S_1、S_2、およびS_3 に関連する出力分布は、それぞれ$b_1(x)$、$b_2(x)$、および$b_3(x)$である。状態S_0 およびS_4 はヌル状態であり、従って関連する出力を有さない。状態S_0 とS_4 を結合することによって等価なHMMが形成されることがかのうである。しかし、HMMを結合してより大きなHMMネットワークを形成するタスクを簡素化するために、このことは行われない。これについては以下に説明が行われる。HMMに関するより深い検討は、Rabiner による「A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition」(Proc.IEEE,vol.77,No.2,February,1989,pp.257-285)に見出される。

6

【００１９】対象物のシーケンスをモデル化するネットワークHMMは、以下のように個々のHMMを並列に結合することにより作成される。認識されるL個の対象物の各々に対するHMMを$L_i, i=1,\ldots,M$とする。先に述べたように、対象物は単語、単音、またはスピーカーのいづれであってもよい。ネットワークHMMは、許容される全ての対象物シーケンスに対して対象物HMM間の遷移を付加することにより作成される。図４において、HMM　L_1、L_2、およびL_3 によって３つの対象物がモデル化されている。これら対象物は、遷移により示されるように任意の順序で発生可能である。状態S_0 はヌル状態であり、従って出力を発生しない。S_0 からは、対象物HMM　L_1、L_2、およびL_3 への遷移は等確率となる。全ての対象物HMMからの遷移は状態S_R に向かい、次に状態S_0 への遷移となる。

【００２０】T個の出力$X = x_1 \ldots x_T$ のシーケンスが与えられる場合、どの対象物HMMシーケンスが最も確からしく出力シーケンスXを発生したかを決定することにより認識が実行される。これにはビタビアルゴリズムが使用され、最も確からしく出力Xを発生したネットワークを介する状態シーケンスが見出される。シーケンス内の各状態は、認識される対象物の内の１つのHMMに対して特定されるため、最も確からしい状態シーケンスは認識対象物のシーケンスを特定する。図５はビタビアルゴリズムの結果を概略的に示す。x軸は時間を示し、y軸はネットワークHMM内の現行状態を示す。HMM　L_1、L_2、およびL_3 に対応する状態はy軸上の領域によって示される。与えられた出力を結果としてもたらし得る状態シーケンスが多数存在可能であるが、ビタビアルゴリズムは最も確からしい状態シーケンスを見出す。図５はビタビパスを示す。時刻t_0 において最も確からしい対象物はL_1 である。時刻t_1 において対象物はL_2 であり、t_2 においてはL_3 である。時刻t_3 において最も確からしい対象物はL_1 となる。

【００２１】HMMに対するパラメータは、次に、遷移確率a_{ij}および出力確率$b_i(x)$である。これらパラメータは、HMMによってモデル化された対象物によって既に発生されたことがわかっている出力Xを用いてHMMをトレーニングすることにより学習されることが可能

である。Baum-Welchプロシジャーとして知られているアルゴリズムが一般に使用される。このアルゴリズムは、トレーニングデータＸの尤度を最大にするパラメータ値を繰り返し処理により見出すアルゴリズムである。該アルゴリズムは、パラメータの初期推定から開始する。続いて以下のステップが実行される。（１）トレーニングデータに基づき、状態間遷移確率および状態からの出力確率を算出する。（２）これらの確率を使用し、遷移確率a_{ij}および出力確率$b_i(x)$の推定値を算出する。ステップ（１）および（２）は収束が得られるまで繰り返される。

【００２２】前述のように、隠れマルコフモデルが使用されてスピーカー識別を目的として個々のスピーカーがモデル化されることが可能である。図６に示されるように、（特定の発声に対向する）個々の発声スタイルが３５状態ＨＭＭ６０を使用してモデル化されることが可能である。状態S_0はヌル状態であり、出力を発生する状態$S_1,\ldots,S_{32}$およびS_{SIL}への遷移を伴う。これらの遷移確率は$p_1,\ldots,p_{32}$およびp_{SIL}により与えられる。これら出力発生状態の各々は、確率q_iを伴う自己遷移、ならびに確率$1-q_i$を伴う最終ヌル状態S_{34}への遷移を有している。ヌル状態S_{34}は確率１で初期ヌル状態S_0へ遷移する。各非ヌル状態はガウシアン出力分布を有しており、平均ベクトルおよび対角共分散マトリックスにより特性付けられる。

【００２３】図７はサイレンス（無音）サブネットワークを示す。該サブネットワークは直列に接続された３状態から成る。各状態は、通常もしくは結合されたガウシアン出力分布を有し、該分布はラベルＳＩＬで示されている。この出力分布はまた、スピーカーモデル６０のサイレンス状態６２における出力分布と同一であり、該分布は状態ラベルＳＩＬで示されている。サイレンスサブネットワークは長時間間隔の無音状態をモデル化するが、会話の発声におけるポーズや短時間間隔の無音状態に対しては適切でない。これらポーズや短時間間隔の無音状態は、個々のスピーカーモデルにおけるサイレンス状態６２によってモデル化される。スピーカーＨＭＭのサイレンス状態における出力分布は全て結合されてサイレンスサブネットワークにおける出力分布となる。

【００２４】スピーカーＨＭＭの各々は、与えられたスピーカーの発声スタイルに対してトレーニングされなければならない。このトレーニングは先に述べたBaum-Welchアルゴリズムを使用して行われ、遷移確率a_{ij}、およびガウシアン出力確率$b_i(x)$に対する平均および対角共分散が推定される。ＨＭＭパラメータの初期推定値は次のように得られる。全ての遷移確率が画一的に設定され、この結果、与えられた状態からの全ての遷移は等確率となる。ガウシアン出力分布を初期化するために、スピーカーに対するトレーニングデータから全体平均および対角共分散マトリックスが算出される。全ての状態に対するガウシアン出力分布についての共分散マトリックスが全体的共分散マトリックスに設定される。全体平均に小さな定数を加えることによって平均が設定され、その場合該定数は異なる各状態に対するランダム要素に対して加えられる。Baum-Welch繰り返し処理がスピーカーのトレーニングデータを用いて次に実行される。

【００２５】認識されるスピーカーが事前にわかっている場合、Baum-Welchアルゴリズムに対するトレーニングデータは、３０秒から１分の各スピーカーに対する音声データを使用して得られる。音声はスピーカーの通常の発声スタイルを表さなければならないが、この場合使用される実際の単語は重要でない。

【００２６】スピーカーおよびサイレンスサブネットワークに加えて、ガーベッジ（garbage)サブネットワークが頻繁に使用され、スピーカーモデルまたは存在可能な非音声音の内の１つによって特定されない任意のスピーカーがモデル化される。ガーベッジネットワークの形態は、図６に示されるスピーカーネットワークのそれと同じである。しかし、アプリケーションに依存してガーベッジネットワークは異なるデータを使用してトレーニングされる。例えば、ガーベッジサブネットワークが使用されて非音声音がモデル化される場合、それはスピーカーモデルとしてトレーニングされなければならないが、この場合非音声データが使用される。システムに対して未知のスピーカーをモデル化する場合、トレーニングデータを得る１つの方法は、既知の各スピーカーからの音声の部分を使用することである。

【００２７】ガーベッジモデルをトレーニングする際に全てのスピーカーからの全てのデータが必ずしも使用されないことは重要である。全ての有効なデータを使用することは、各スピーカーモデルに対してよりもガーベッジモデルに対してより多くのトレーニングデータを与え、全てのスピーカーに対してより確実なスピーカーモデルを作成する効果を有する。従って、結果として得られるＨＭＭネットワークはほとんどの音声をガーベッジとして分類する。

【００２８】１実施例において、入力オーディオトレーニングデータは８ＫＨｚでサンプルされ、１０ｍｓごとに特徴ベクトルが算出される。例えば、各フレームに対する特徴ベクトルは、２５ｍｓウィンドウ下のサンプルに関する２０次の線型予測符号化（ＬＰＣ）を行うことによって算出されることが可能であり、従ってＬＰＣスペクトルから２０個のケプストラム定数が算出されることが可能である。

【００２９】いくつかの場合においては、認識されるスピーカーは事前にわかっていない。しかし、スピーカーモデルに対する初期推定を得ることがそのような場合にも必要である。この初期推定は、階層的な集塊性のクラスタリングを使用して行われ、異なるスピーカーとして認識されるデータのラフな区分が作成される。

9

【００３０】スピーカーに従うデータの区分を与えることによってスピーカーサブネットワークの初期推定を得るために、階層的な集塊性のクラスタリングが使用されることが可能である。このデータは次にスピーカーＨＭＭのBaum-Welchトレーニングに対するトレーニングデータとして使用されることが可能である。

【００３１】セグメンテーションされていないデータは、最初に等しい長さのセグメントに分割され、各セグメントは数秒の音声から成る。これらのセグメントは階層的クラスタリングに対する初期クラスタ集合として使用される。該アルゴリズムは、最初に全てのクラスタペアについてのクラスタ間距離を算出し、次に最も近い２つのクラスタを併合することによって進行する。このプロセスは所望のスピーカークラスタ数が得られるまで繰り返される。このプロセスが図８に概略的に示されている。スピーカー数が未知の場合、このアルゴリズムが使用されてスピーカー数が推定されることが可能である。その場合、最近接クラスタの併合は、最近接クラスタ間距離が定められたスレショルドを越えるまで継続する。スレショルドを越えるとクラスタリングは中止され、その時のクラスタ数がスピーカー数の推定値として使用される。

【００３２】図８は、スピーカーでラベル付けされているインターバル集合上の階層的クラスタリング１００を概略的に示す。オリジナルインターバル１０２は、Ｃ、Ｌ、およびＴで３つのスピーカーに対してラベル付けされたツリーのリーフによって示される。そのような全てのインターバルについてのインターバル間距離が算出され、１０４に示されるように最も近接する２つのインターバルが併合される。

【００３３】この最近接クラスタ併合プロセスは、所望のクラスタ数が形成されるまで繰り返される。３つのクラスタに対し、それらクラスタに対応する３つの分岐が示されている。第１のクラスタ１０６はほとんどスピーカーＣからのインターバルを含み、第２のクラスタ１０８はほとんどスピーカーＬからのインターバルを含み、第３のクラスタ１１０はほとんどスピーカーＴからのインターバルを含む。

【００３４】スピーカー数が未知の場合、距離に対するスレショルドが設定され、スレショルドが越えられた場合にクラスタの併合が中止される。このことは線１１２により概略的に示されており、該線は４つのクラスタを生成する。（クラスタ１は２つに分割されている。）クラスタＸが単一セグメントＸ＝ｘかまたはセグメント集合Ｘ＝$x_1, x_2, \ldots$から成ると仮定する。クラスタＸおよびＹ間の距離はｄ（Ｘ，Ｙ）により表される。前述のシステムにおいて、セグメント間距離はガウシアン分布の仮定に基づき尤度比によって導出された。ｘ＝$s_1, \ldots, s_r$はある１つのセグメント内のデータを表し、ｙ＝$s_{r+1}, \ldots, s_n$はその他のセグメント内のデータを表し、ｚ＝$s_1, \ldots, s_n$は合成セグメント内のデータを表すものとする。Ｌ（ｘ，θ_x）はｘシーケンスの尤度とし、ここでθ_xはガウシアン分布のパラメータに対する推定値である。同様にＬ（ｙ，θ_y）はｙシーケンスの尤度とし、Ｌ（ｚ，θ_z）は合成シーケンスｚの尤度とする。λは尤度比を表すとすると、次式のように表される。

10

【００３５】

【数１】

$$\lambda = \frac{L(z, \theta_z)}{L(x, \theta_x)\, L(y, \theta_y)}$$

【００３６】クラスタリングの際に使用される距離計量は－ｌｏｇ（λ）である。音声データは単一のガウシアン分布では充分にモデル化されないため、尤度比はガウシアン分布の混成結合に拡張される。セグメンテーションされていないデータが最初に使用され、Ｍ個のガウシアン分布の混成に対する平均および共分散マトリックスが推定される。次にこれらは残りの解析により確定される。N_i（ｓ）＝Ｎ（ｓ：M_i，σ_i）はｉ番目の混成要素に関連するガウシアン分布とし、g_i（ｘ）はデータシーケンスｘを使用して推定されたｉ番目の混成要素に対する重みとする。g_i（ｘ）はN_i（ｓ）が最大となるｘ内のサンプルの割合である。従ってｘシーケンスの尤度は次式のように表される。

【００３７】

【数２】

$$L(x, \theta_x) = \prod_{j=1}^{r} \sum_{i=1}^{M} g_i(x)\, N_i(s_j)$$

【００３８】ここでθ_x＝（g_1（ｘ），…，g_M（ｘ））である。尤度Ｌ（ｙ，θ_y）も同様に算出される。合成シーケンスに対する尤度Ｌ（ｚ，θ_z）の算出において、混成要素に対する重みg_i（ｚ）として次式を得る。

【００３９】

【数３】

$$g_i(z) = (r/n)\, g_i(x) + ((n-r)/n)\, g_i(y)$$

【００４０】クラスタリングに対する距離計量、d_L＝－ｌｏｇ（λ_L）は従って式（１）を使用して算出されることが可能である。

【００４１】本発明のクラスタリングプロシジャーは、クラスタを含むインターバルにおけるインターバル間距離の最大、最小、もしくは平均を使用するよりもむしろ式（１）を使用して集塊性のクラスタ間距離を再計算する点において、通常の階層的クラスタリングと異なっている。従って式（２）および（３）により与えられる尤度の計算効率が重要となる。これはクラスタリングレベルの各々において距離が再計算されるためである。

11

【００４２】加えて、スピーカーチェンジの事前確率はＭ個のスピーカーを伴うマルコフデュレーションモデルを使用して算出されることが可能である。S_i はセグメントｉの期間中のスピーカーを表し、Ｍはスピーカー数を表すとする。S_i は、各スピーカーａに対してP_r〔S_{i+1} ＝ａ｜S_i ＝ａ〕＝ｐ、および各スピーカーａおよびｂ（ａに等しくない）に対してP_r〔S_{i+1} ＝ｂ｜S_i ＝ａ〕＝（１－ｐ）／（Ｍ－１）を伴うマルコフ連鎖であると仮定する。セグメントｉに対するスピーカーがセグメントｉ＋ｎに対しても発声する確率P_r〔S_{i+n} ＝S_i〕は、２状態マルコフ連鎖を使用して算出されることが可能であり、その場合連鎖の状態１は時刻ｉにおけるスピーカーを表し、状態２は他の全てのスピーカーを表す。この連鎖に対する遷移確率マトリックスＰは次式のように表される。

【００４３】

【数４】

$$\mathbf{P}=\begin{pmatrix} p & 1-p \\ \dfrac{(1-p)}{M-1} & 1-\dfrac{(1-p)}{M-1} \end{pmatrix}$$

【００４４】このマトリックスに関し、P_r〔S_{i+n} ＝S_i〕＝$(P^n)_{11}$である。Ｐを対角化することにより、P_r〔S_{i+n} ＝S_i〕は次式のようによりクローズした形態で表されることが可能である。

【００４５】

【数５】

$$f(n)=\Pr\{S_{i+n}=S_i\}=\frac{1+(M-1)\left(\frac{Mp-1}{M-1}\right)^n}{M}$$

【００４６】この式を使用して、２つの与えられたクラスタが同一のスピーカーまたは２つの異なるスピーカーによって生成される事前確率を算出することが可能である。Ｃをスピーカーチェンジが発生するインターバル数とし、n_i をｉ番目のインターバル長とすると、デュレーションバイアスは次式のように定義される。

【００４７】

【数６】

$$\lambda_D=\frac{\prod_i^C f(n_i)}{(M-1)\prod_i^C (1-f(n_i))/(M-1)}$$

【００４８】デュレーションバイアスされた距離はd_D（Ｘ，Ｙ）＝－ｌｏｇ（λ_L）－ｌｏｇ（λ_D）として定義される。

【００４９】図９に示されるスピーカーセグメンテーションネットワーク１２０は、各スピーカーに対するサブネットワーク６０と、サイレンスおよびガーベッジに対するオプショナルなサブネットワーク６４および１２２とから成る。ガーベッジは、オーディオ中の未知のスピ

12

ーカーまたは非音声音のような、スピーカーまたはサイレンスモデルによってモデル化されない音声または音として定義される。スピーカー、ガーベッジ、およびサイレンスサブネットワークは以下に述べるように得られる。ネットワークモデルは、２またはそれ以上のスピーカーによるバックグランドノイズを伴う会話をモデル化する。

【００５０】ネットワーク６０のような個々のスピーカーサブネットワークは互いに並列に結合され、各サブネットワークから外部への遷移確率は小さいペナルティ定数εに固定されて、孤立サンプルに基づくスピーカーチェンジが抑制される。各スピーカーサブネットワーク６０はＬ個の状態を伴うＨＭＭから成り、それらＨＭＭは並列に接続される。各状態は、ガウシアン出力分布、自己遷移、および他状態への遷移を有する。

【００５１】初期ヌル状態からスピーカー、ガーベッジ、およびサイレンスサブネットワークへの遷移確率は画一的である。スピーカー、ガーベッジ、およびサイレンスモデルから外部への遷移確率ペナルティは定数εに設定される。原理的に、これら遷移確率はスピーカーに依存し、トレーニング期間中に学習される。しかし、簡素化を目的として、スピーカーの事前確率は画一値に仮定され、スピーカーを離れる確率εは経験的に選択されて孤立サンプルに基づくスピーカーチェンジが抑制される。

【００５２】実際には、この遷移確率は著しく小さい。（10^{-20} のオーダーである。）従って各スピーカーモデルから外部への遷移は、スピーカーからスピーカーへの切替にペナルティを与えるよう作用する。

【００５３】スピーカー間の会話をインデックス化することは単に、観測された特徴ベクトルに関する与えられたシーケンスであるネットワークモデルを介する最も確からしい状態シーケンスを見出すことである。スピーカーサブネットワークが初期化された後、スピーカーセグメンテーションネットワークを介する最も確からしい状態シーケンスを見出すことによりスピーカーセグメンテーションが実行され、状態パスがスピーカーを変更する時点でマーキングが施される。最適な状態が１つのスピーカーモデルから他のスピーカーモデルへ切り替わる場合にスピーカーチェンジが発生する。最適な状態シーケンスを見出すことはビタビアルゴリズムを使用して達成される。セグメンテーションの確度は、セグメンテーションされたデータを使用してスピーカーサブネットワークを再トレーニングすることによって改善されることが可能である。このセグメンテーションおよび再トレーニングのプロセスは、セグメンテーションにおいて変化が生じなくなるまで繰り返される。

【００５４】部分的トレースバックの方法または連続的デコーディングがビタビ探索で使用される。部分的トレースバックは、Brown らによる「Partial Traceback an

13

d Dynamic Programming」(Proc.Int.Conf.Acoustics,Speech and Signal Processing,May 1992,pp.1629-1632)に記載されている。このアルゴリズムにおいて、全ての状態からの各タイムステップにおいてビタビトレースバックが実行され、全てのパスの初期部分が整合する場合にはデコーディングが可能となる。実際に被る遅延は１秒未満である。

【００５５】後処理のアプリケーションにおいて、音声のセグメンテーションが繰り返し実行されることが可能であり、その場合各セグメンテーションの後にスピーカーモデルが再トレーニングされる。このことはセグメンテーションの確度を向上させ、特にスピーカートレーニングデータが使用不可能な場合に有効である。

【００５６】繰り返し再セグメンテーションアルゴリズムが図９に示される。前述のように、最初にトレーニングデータ集合がボックス１３０のステップで与えられ、ボックス１３２のステップでスピーカーモデルがトレーニングされる。次にボックス１３４のステップでこれらスピーカーモデルに基づきセグメンテーションが実行される。ボックス１３４のステップでのセグメンテーションが大きく変化する場合、この改善されたセグメンテーションはスピーカーに対する新たなトレーニングデータとして使用され、ボックス１３２のステップでスピーカーモデルが再トレーニングされる。このプロセスはボックス１３６のステップでセグメンテーションが変化しなくなるまで続けられる。

【００５７】図１０は、オーディオデータが記録デバイスによって記憶媒体上に記憶される場合にリアルタイムでオーディオデータのインデックスを作成する、本発明のシステム１４０を示す。

【００５８】メモリ１４８から命令を得るシステムプロセッサ１４６はトレーニングデータ１４７を受信してスピーカーモデルを決定する。スピーカーモデルは結合され、後のオーディオストリーム処理のためのスピーカーネットワークが形成される。トレーニングデータ１４７は、識別される各スピーカーに対するトレーニングデータを有していなければならない。図１０に示されるように、トレーニングデータ１４７はそのオリジナルなオーディオ波形から既に処理されており、スペクトル特徴データの形態でシステムプロセッサ１４６に保存されている。

【００５９】オーディオ入力１４１はオーディオプロセッサ１４２によってスペクトル特徴データへ処理され、システムプロセッサ１４６に与えられる。これと同時に、オーディオ入力はオーディオ記録デバイス１４３によって記憶媒体１４４上に記録される。記録デバイス１４３は、オーディオストリーム情報をアナログまたはデジタル形態で記憶することが可能であり、純粋なオーディオ記録、もしくはオーディオ／ビデオ記録の部分であることが可能である。

14

【００６０】スペクトルデータは、システムプロセッサ１４６によってトレーニングデータ１４７から作成されたスピーカーネットワークを使用することによってシステムプロセッサ１４６によって処理される。オーディオストリームにおいて新たなセグメントの各々が検出されると、システムプロセッサ１４６はタイムソース１４５からタイムスタンプを得る。タイムスタンプは、オーディオデータの記憶媒体１４４上への記憶時間を示す。タイムソース１４５は、例えば、記録が開始される時に始動する時計であることが可能であり、もしくは、記憶媒体に接続された記録デバイスから時間を記録するデバイスであることが可能である。このタイムスタンプは、セグメントの作成者の識別子と共にメモリ１４８に記憶され、後にスピーカーに従うインデックスへ収集される。

【００６１】図１１は、オーディオ記録データのスピーカーに従うインデックスを作成および記憶する、システム１９０における本発明のその他の実施例を示す。

【００６２】トレーニングデータ１９６はシステムプロセッサ１９４へ与えられ、スピーカーモデルおよびスピーカーネットワークが生成される。トレーニングデータ１９６は識別される各スピーカーに対するトレーニングデータを有していなければならない。図１１に示されるように、トレーニングデータは、既にそのオリジナルなオーディオ波形から処理されており、スペクトル特徴データとしてシステムプロセッサ１９４に保存されている。識別された各スピーカーに対して記録の部分が孤立され得る場合、トレーニングデータはまたオーディオ記録入力１９１の部分であることが可能である。

【００６３】オーディオ記録入力１９１はオーディオプロセッサ１９２によってスペクトル特徴データへ処理され、システムプロセッサ１９４へ与えられる。スペクトル特徴データは、システムプロセッサ１９４による後の繰り返し処理のためにメモリ１９７に記憶されることが可能である。スペクトルデータは、システムプロセッサ１９４によってトレーニングデータ１９６から作成されたスピーカーネットワークを使用することによってシステムプロセッサ１９４によって処理される。オーディオストリームにおいて新たなセグメントの各々が検出されると、システムプロセッサ１９４はタイムソース１９３からタイムスタンプを得る。タイムスタンプは、オーディオ入力１９１の記録からのオーディオデータの記録アドレスまたは記憶時間を示す。タイムソース１９３は、例えば、記録が開始される時に始動する時計であることが可能であり、もしくは、記憶媒体に接続された記録デバイスから時間を記録するデバイスであることが可能である。このタイムスタンプは、セグメントの作成者の識別子と共にメモリ１９５に記憶され、後にスピーカーに従うインデックスへ収集される。

【００６４】図１０のシステム１４０によって記録されるオーディオデータは、図１２のシステム１９０におい

て記録データ１９１として使用されることが可能である。そのような場合、システム１４０によって作成されるインデックスが使用されてトレーニングデータ１９６が与えられることが可能であり、特定のスピーカーに属するセグメントの各集合は、新たなスピーカーモデルをトレーニングするトレーニングデータとして使用される。システムプロセッサ１９４は新たなスピーカーモデルを使用し、それらを結合してネットワークとし、オーディオストリームの再セグメンテーションを行う。

【００６５】そのような繰り返し処理は、システム１４０からシステム１９０へのものであれ、もしくはシステム１９０を繰り返し介するものであれ、セグメンテーションの確度をさらに向上させる。

【００６６】図１２は、オーディオストリームのインデックスを決定する前述の方法のアプリケーションを示す。図１２におけるステップはリアルタイムもしくは後処理モードで実行されることが可能である。オーディオストリームは通常、オーディオタイミングに相関付けられたセグメント情報と共に記録される。ボックス１５０のステップは既知のスピーカーからトレーニングデータを選択する。前述のように、そのようなトレーニングデータは個々のスピーカーによる３０秒から１分の音声から成ることが可能である。このトレーニングデータがボックス１５２のステップで使用されて個々のスピーカー各々に対するＨＭＭスピーカーモデルがトレーニングされる。

【００６７】ボックス１５４のステップにおいて、図８に関連して述べられたように、個々のモデルが並列に接続されてスピーカーセグメンテーションネットワークが形成される。この時点で、個々のスピーカーモデルから離れることに対するペナルティが挿入される。ボックス１５６のステップはガーベッジ、即ち未知のスピーカーおよび／または非音声音、およびサイレンスインターバルに対するモデルを作成および付加する。サイレンスおよびガーベッジモデルはボックス１５２のステップで既に作成されていることも可能である。

【００６８】ボックス１５８のステップにおいて、オーディオストリームはスピーカーセグメンテーションネットワークを使用してセグメントに分解される。セグメントは、ボックス１６０のステップにおいて各セグメントに対するスピーカーの識別子を用いてマーキングされる。ボックス１６２のステップは同様のマーキングが施されたセグメントを収集してオーディオ記録のスピーカーインデックスを作成する。

【００６９】リアルタイム動作が必要ない場合、図１３に関して述べられるように、より詳細な処理が実行されることが可能である。ボックス１７０から１８０に示されるステップは、図１２のボックス１５０から１６０に示されるステップに関して述べられた方法と同様に実行される。

【００７０】ボックス１８２に示されるステップにおいて、テストが実行され、ボックス１７８のステップで決定されたセグメンテーションが前のセグメンテーションから変化したかどうかが決定される。セグメンテーションに大きな変化があった場合、システムはボックス１７２のステップに戻り、スピーカーモデルの再トレーニングおよびオーディオストリームの再セグメンテーションが行われる。シーケンスの初期には前のセグメンテーションが存在しないためシステムは前述のように繰り返しを行う。オーディオストリームのセグメンテーションにおいて繰り返し処理により大きな変化が生じなくなった場合、同様のマーキングが施されたセグメントが収集されてボックス１８４のステップでインデックスが作成されることが可能である。

【００７１】情報のインデックス化は、スピーカーＩＤに基づきオーディオストリームの再生を検索および制御する能力をユーザーに与えることが可能である。例えば、ユーザーは特定のスピーカーによる発言のみを検索したいかもしれない。ユーザーはさらに、オーディオインデックス情報を使用してオーディオ記録を検索するかもしれない。ユーザーは、いくつかのスピーカーセグメントをスキップし、次のスピーカーへ効果的に早送りを行うかもしくは特定のスピーカーセグメントの始まりへ巻戻しを行いたいかもしれない。

【００７２】

【発明の効果】以上説明したように、本発明の方法によれば、オーディオストリーム内にインデックスを作成することが可能となり、リアルタイムであっても処理後であっても、ユーザーが特定のスピーカーに関連するオーディオデータセグメントを認識することが可能となる。

【図面の簡単な説明】

【図１】本発明が実施されることが可能である一般化されたオーディオ処理システムのブロック図である。

【図２】オーディオインデックスシステムの一般化されたフロー図である。

【図３】５状態隠れマルコフモデル（ＨＭＭ）を示す図である。

【図４】ＨＭＭによってモデル化される３つの対象物のＨＭＭネットワークを示す図である。

【図５】ビタビアルゴリズムの結果を概略的に示す図である。

【図６】個々のスピーカーの発声スタイルをモデル化する３５状態ＨＭＭを示す図である。

【図７】サイレンスサブネットワークを示す図である。

【図８】各スピーカーに対するサブネットワークと、サイレンスおよびガーベッジに対するオプショナルなサブネットワークとから成るスピーカーセグメンテーションネットワークを示す図である。

【図９】繰り返し再セグメンテーションアルゴリズムを概略的に示す図である。

17

【図１０】オーディオデータが記録デバイスによって記憶媒体上へ記憶される場合にリアルタイムにオーディオデータのインデックスを作成する本発明に従うシステムを示す図である。

【図１１】事前に記録されたオーディオデータのスピーカーに従うインデックスを作成および記憶するシステムにおける本発明のその他の実施例を示す図である。

【図１２】スピーカーが未知の場合にオーディオストリームのインデックスを決定する本発明に従う方法を示す図である。

18

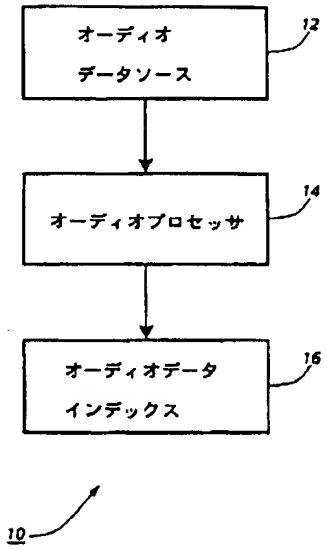
【図１３】オーディオストリームのインデックスを決定する際に実行されることが可能なより詳細な処理を示す図である。

【符号の説明】

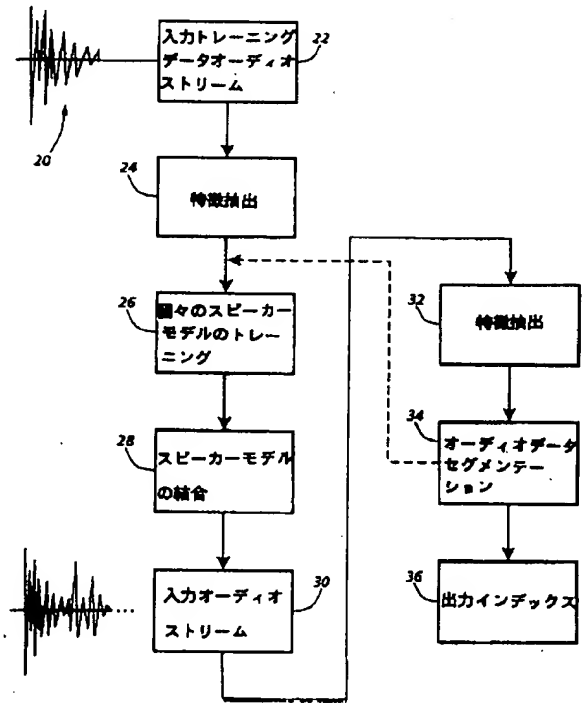
１２　オーディオデータソース
１４　オーディオプロセッサ
１６　オーディオデータインデックス
６０　３５状態ＨＭＭ
１２０　スピーカーセグメンテーションネットワーク

10

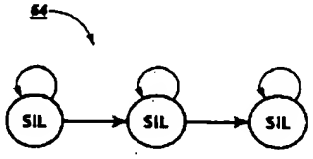
【図１】



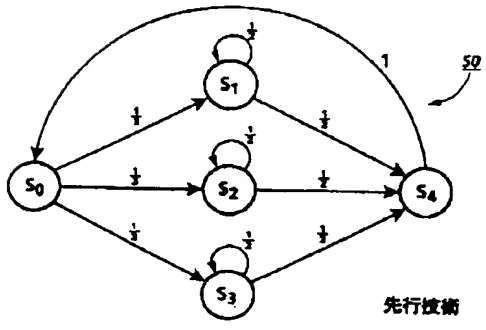
【図２】



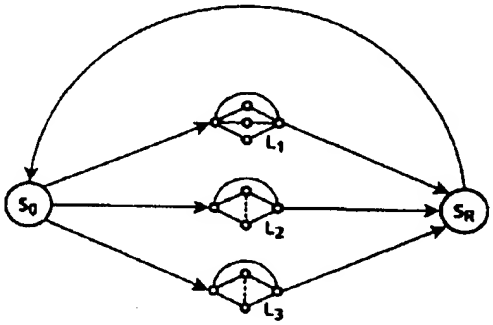
【図７】



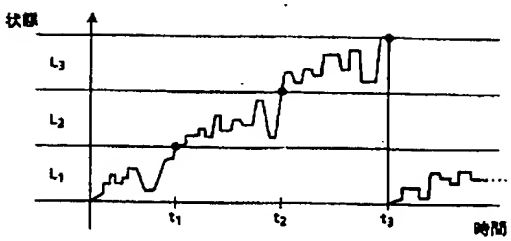
【図３】



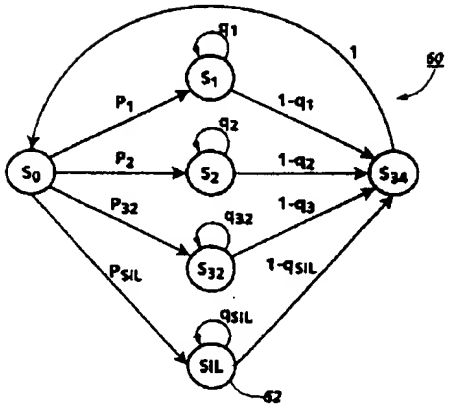
【図４】



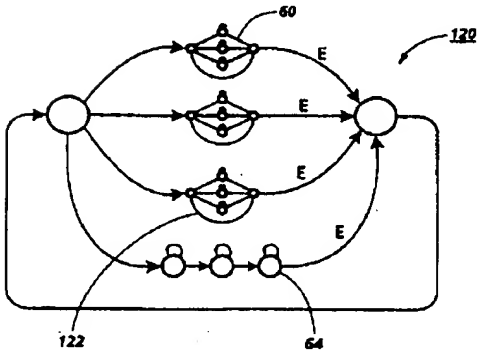
【図5】



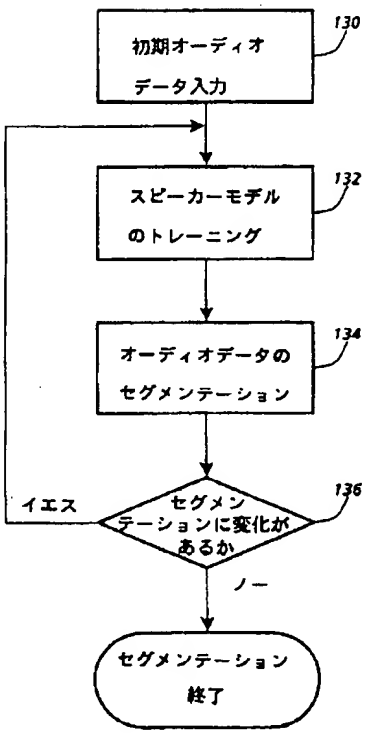
【図6】



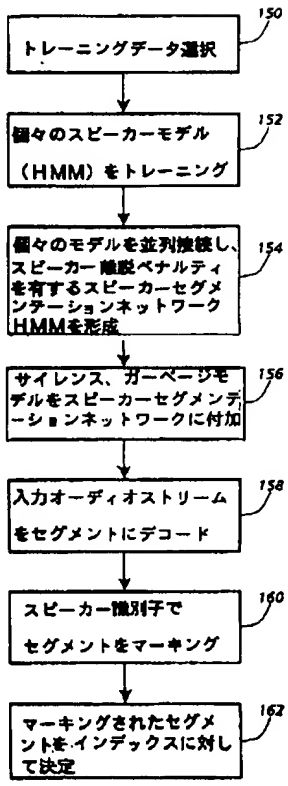
【図8】



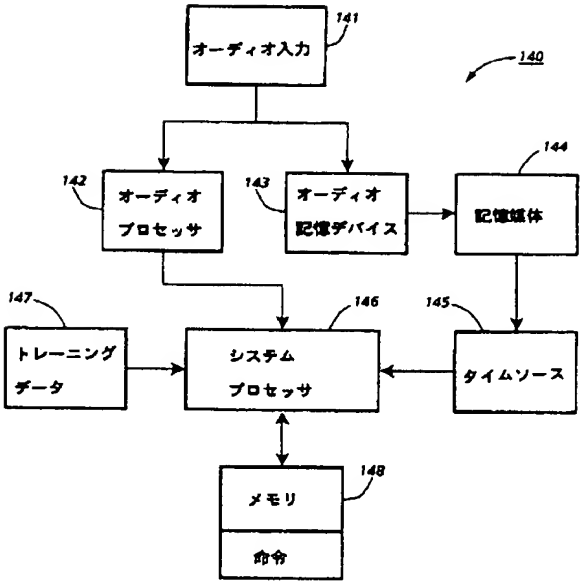
【図9】



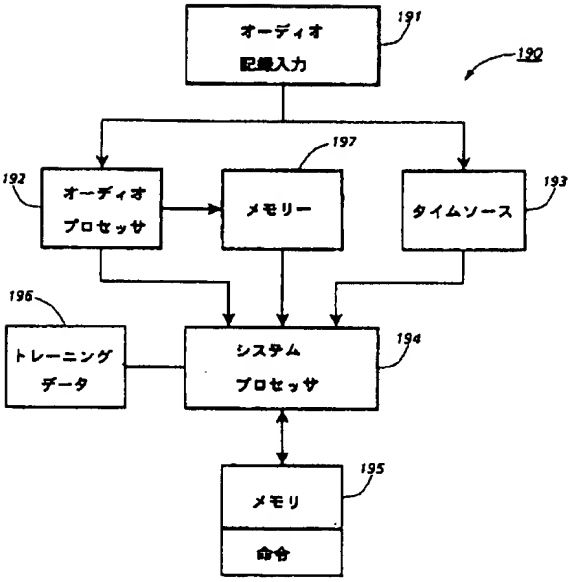
【図12】



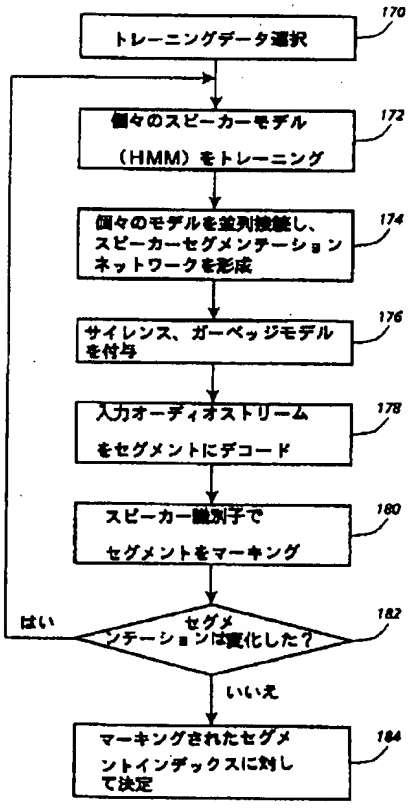
【図１０】


オーディオ入力
141
140
142
オーディオ
プロセッサ
143
オーディオ
記憶デバイス
144
記憶媒体
147
トレーニング
データ
146
システム
プロセッサ
145
タイムソース
148
メモリ
命令

【図１１】


オーディオ
記憶入力
191
190
197
192
オーディオ
プロセッサ
メモリー
タイムソース
193
196
トレーニング
データ
システム
プロセッサ
194
メモリ
195
命令

【図１３】


トレーニングデータ選択
170
個々のスピーカーモデル
（HMM）をトレーニング
172
個々のモデルを並列接続し、
スピーカーセグメンテーション
ネットワークを形成
174
サイレンス、ガーベッジモデル
を付与
176
入力オーディオストリーム
をセグメントにデコード
178
スピーカー識別子で
セグメントをマーキング
180
はい
セグメ
ンテーションは変化した？
182
いいえ
マーキングされたセグメ
ントインデックスに対し
て決定
184

フロントページの続き

(72)発明者　フランシン　アール．チェン
アメリカ合衆国　カリフォルニア州　94025　メンロ　パーク　シャーマン　アヴェニュー　975
(72)発明者　フィリップ　エイ．チョウ
アメリカ合衆国　カリフォルニア州　94025　メンロ　パーク　ブラックバーン　アヴェニュー　116
(72)発明者　ドナルド　ジー．キンバー
アメリカ合衆国　カリフォルニア州　94040　マウント　ビュー　ヴィクター　ストリート　678　ナンバー　３
(72)発明者　アレックス　ディー．プーン
アメリカ合衆国　カリフォルニア州　94040　マウンテン　ビュー　サウス　レングストーフ　アヴェニュー　575　アパートメント　ナンバー　21
(72)発明者　カロン　エイ．ウェバー
アメリカ合衆国　カリフォルニア州　94109　サンフランシスコ　ユニオン　ストリート　1330　ナンバー　22
(72)発明者　リン　ディー．ウィルコックス
アメリカ合衆国　カリフォルニア州　94028　ポートラ　ヴァレー　ジョアクイン　ロード　45